DOCUMENT RESUME

ED 317 597 TM 014 690

AUTHOR Kim, Gyenam; Edwards, Lynne K.
TITLE Multiple Comparisons of Means with an Emphasis on Recent Developments.
PUB DATE Apr 90
NOTE 28p.; Paper presented at the Annual Meeting of the American Educational Research Association (Boston, MA, April 16-20, 1990).
PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Comparative Analysis; Equations (Mathematics); *Mathematical Models; Research Methodology; *Statistical Analysis
IDENTIFIERS Bonferroni Procedure; Dunnetts Test; *Multiple Comparisons of Means; Power (Statistics); Scheffes Contrast Test; *Sequentially Rejective Bonferroni Test; Tukeys Test

ABSTRACT

An integrated overview is provided of selected multiple comparison methods, with an emphasis on the sequentially rejective Bonferroni (SRB) test and its modifications. Multiple comparisons of means is a frequently used, and yet confusing, method in educational and psychological research. Seven multiple comparison methods are reviewed: (1) the classical Bonferroni t test (CB); (2) Tukey's HSD test (1953); (3) H. Scheffe's test (1953); (4) C. W. Dunnett's test (1955); (5) the SRB; (6) a modification of the SRB (MSRB1); and (7) another modification of the SRB (MSRB2). It is concluded that the SRB introduced by S. Holm (1979) is flexible and easy to use in various research situations and has more power than does the CB procedure. The modified SRB procedures described give more power than does the simple SRB, although they have restricted applications. If a researcher prefers a simple procedure with reasonable power in conducting all pairwise comparisons, either the Tukey test or the SRB seems to be the best choice. All of the described procedures control experimentwise error rates at the nominal level of significance. The choice among these procedures depends on the experimenter's decisions about the type of contrast that is the primary interest, the most powerful procedure, and the desired computational simplicity. Some of the recent multiple comparison procedures are presented geometrically to aid visual understanding of these complex methods. Eight tables and seven figures are included. (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED317597

Multiple Comparisons of Means with an Emphasis on Recent Developments

Paper presented at the Annual Meeting of the American Educational Research Association in Boston (April 21, 1990).

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
LYNNE K. EDWARDS
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Gyenam Kim and **Lynne K. Edwards**

Department of Educational Psychology
University of Minnesota

Computer funds from the Minnesota Supercomputer Institute are gratefully acknowledged. Requests for reprints should be sent to Lynne K. Edwards, Dept. of Educational Psychology, 323 Burton Hall, 178 Pillsbury Dr. S.E., Minneapolis, MN 55455.

TM014690

Introduction

Multiple comparisons of means is one of the most frequently used and yet most confusing methods in educational and psychological research. Games (1971) provided an excellent review of the available multiple comparison procedures and Jaccard, Becker, and Wood (1984) reviewed the conventional procedures for pairwise comparisons of means for researchers in behavioral sciences. A new method, called the sequentially rejective Bonferroni (SRB) test, was introduced by Holm (1979). This new method is widely applicable to pairwise and various contrasts with fewer restrictions and is more powerful than the conventional methods. The SRB has been further modified by Shaffer (1986) and Holland and Copenhaver (1987).

The purpose of this paper is to provide an integrated overview of the selected multiple comparison methods with an emphasis on the SRB and its modifications, because it is still confusing as to when these procedures are applicable and appropriate, even though research efforts related to the SRB are accumulating in recent years. It is hoped that this study can provide a practical guidance for researchers in choosing appropriate multiple comparison procedures.

Seven multiple comparison methods: classical Bonferroni t test (CB), Tukey's HSD, Scheffé's test, Dunnett's test, the sequentially rejective Bonferroni test (SRB), and two types of the modified sequentially rejective Bonferroni tests (MSRB1 and MSRB2) are examined in this paper.

Classical Bonferroni t procedure (CB)

Dunn (1961) introduced a multiple comparison procedure on means using the Bonferroni inequality, which is a general method of allocating α for testing hypotheses so that the experimentwise type I error rate does not exceed the specified level. The multiple comparison procedure proposed by Dunn is appropriate when the total number of contrasts to be tested is known. Tests involving these planned hypotheses may not be necessarily

orthogonal.

A null hypothesis for the i^{th} linear combination of k means $(H_o: \Psi_i = o)$, $\Psi_i = c_1\mu_1 + c_2\mu_2 + \ldots + c_k\mu_k$, can be tested by

$$t = \frac{\hat{\Psi}_i}{\sqrt{MSe\sum_{j=1}^{k}\frac{c_j^2}{n_j}}},$$

where $\hat{\Psi}_i$ is the unbiased estimate of Ψ_i and MSe is the pooled estimate of the population error variance. The above t statistic, sometime referred to as the CB, is distributed as Student's t with ν degrees of freedom associated with the MSe. The critical values corresponding to each contrast are not typically reported in standard t-tables. Two special tables by Bailey (1977) and Dunn (1961) provide exact critical values depending on a number of contrasts and degrees of freedom, ν, associated with the MSe. A null hypothesis is rejected if the obtained value of the CB is equal to or greater than the critical value.

Dunn has compared the length of confidence intervals between her procedure and Scheffé s method. Confidence intervals for the CB are shorter than those for Scheffé's if there are a large number of means but the number of contrasts is small. That is, as the degrees of freedom for the MSe increase and if $k \leq 10$, the CB becomes more powerful than Scheffé's test in detecting possible differences. From the comparison of the lengths of confidence intervals for linear contrasts between the CB and the Tukey HSD, the CB intervals are shorter than the HSD if the number of contrasts made are relatively small in number (Dunn, 1961, p.61). In general, the CB yields more powerful results than Tukey's HSD or Scheffé's test under the conditions where there are a large number of means but only a few contrasts are made.

In conclusion, if the contrasts among means are planned in advance and are relatively limited in number but not necessarily orthogonal, the CB can be a powerful procedure to test hypotheses for linear combinations of means.

Tukey 's procedure

Tukey (1953) devised a procedure for testing all pairwise comparisons based on the sampling distribution of q, the studentized range statistic (Miller, 1981). The studentized range statistic for testing the overall range hypothesis $H_o: \mu_1 = \cdots = \mu_k$, or equivalently, $H_o: \mu_{largest} - \mu_{smallest} = 0$, is represented by:

$$q = \frac{\bar{Y}_{largest} - \bar{Y}_{smallest}}{\sqrt{\frac{MSe}{n}}}.$$

There exists a unique relationship between t and q, such that $q=\sqrt{2}\,t$, for their statistics and critical values. The critical value of the q statistic, denoted by $q_{\alpha;k,\nu}$, indicates the upper 100α percentile point on the studentized range distribution with two parameters, k and ν. By using the largest range test as a first step, the type I error is controlled at the nominal level of significance for a set of all possible pairwise comparisons. If q obtained is larger than the critical value, it is concluded that the overall null hypothesis over the range of k means is rejected.

It has been suggested that the Tukey test can be used after the overall F-test is rejected (Hopkins & Chadbourn, 1967; Kirk, 1982; Scheffé, 1959). However, Howell and Games (1973) have indicated that a significant F-test is not necessary in using the studentized range statistic.

When the largest range hypothesis is rejected at α level of significance, Tukey's procedure is commonly used to test all possible pairwise comparisons as a post hoc procedure. Tukey's studentized range statistic for testing, $H_o: \mu_i = \mu_j$, $i \neq j$, is

$$q = \frac{\bar{Y}_i - \bar{Y}_j}{\sqrt{\frac{MSe}{n}}}.$$

The q statistic requires more strict assumption than the CB. The q requires equal sample sizes and homogenous variances across k groups, though there are modifications for dealing with unequal sample sizes and/or variances. Because the tables of the studentized

range statistic are prepared for equal sample sizes, the tables for critical values are not readily available for unequal sample sizes.

There are two different critical values for the q statistic: the HSD and the WSD. Tukey's HSD procedure uses the critical value for the largest range for any pairwise comparison; and the WSD uses the average of the two critical values: one from the Tukey HSD and the other from the Newman-Keuls test (Newman, 1939; Keuls, 1952). The critical values for the Tukey HSD and the Tukey WSD are exactly the same when the largest range is tested. However, the critical values of the WSD become smaller as the steps between two ordered means decrease. Subsequently, the WSD is potentially more powerful than the HSD after the largest range is judged to be significant.

It is possible to extend Tukey's procedures beyond pairwise comparisons to linear contrasts or to linear comparisons. However, it has been shown that the HSD is more powerful than the CB or Scheffé test for making all pairwise comparisons but the HSD loses its power advantage in making nonpairwise linear contrasts (Scheffé, 1953; Dunn, 1961).

Scheffé 's procedure

The Scheffé's statistic for testing a hypothesis, $H_o: \psi_i = 0$, among k means using the t-distribution is:

$$t = \frac{\hat{\psi}_i}{s_{\psi_i}},$$

where $\hat{\psi}_i$ is an estimate of the i^{th} linear combination of means, ψ_i, and s_{ψ_i} is an estimate of the standard error of ψ_i. The critical value of Scheffé's method is

$$S = \sqrt{(k-1)F_\alpha(k-1, \nu)},$$

where $F_\alpha(k-1, \nu)$ is the upper 100α percentile point on the F-distribution with k-1 for the numerator degrees of freedom and ν for the denominator degrees of freedom associated with the MSe.

In Scheffé's procedure, there is a one-to-one relationship between the overall null

hypothesis and the post hoc comparisons among k means. That is, if the overall null hypothesis is not rejected at α, no null hypothesis for any post hoc comparison, $H_o: \psi_i = 0$, will be rejected. On the other hand, if the overall null hypothesis is rejected at α, there exists at least one comparison that will be rejected at α.

In Scheffé's test the value of the contrast obtained is compared against a large critical value. Therefore, a large effect size is required for the hypothesis to be rejected. As the number of contrasts approach to infinity, the type I error rate for a set of contrasts is close to a nominal level of significance (Games, 1971). Subsequently, the error rate per comparison is extremely low. An advantage of Scheffé's procedure is that it can test all possible contrasts that are not limited to pairwise comparisons (Scheffé, 1953; Dunn, 1961). Unlike Tukey's procedure, Scheffé's method can be readily used with the unequal sample sizes. Scheffé's test can be used regardless of how many comparisons are to be made, while maintaining the familywise type I error is controlled at the nominal level of significance (Scheffé, 1953; Dunn, 1961). It is extremely conservative, especially, when the desired number of contrasts is small in relation to k.

The confidence intervals based on the Scheffé test are shorter than that of the CB when the number of groups is relatively small in proportion to the number of contrasts (Dunn, 1961). Even though Scheffé's test can be used as one of the most flexible post hoc procedure, it is not plausible to make infinite number of comparisons which may all result in null findings (Games, 1971). Hence, an experimenter needs to consider employing alternative methods which are more appropriate and more powerful than the Scheffé test in detecting possible differences depending on each experimental situation.

Dunnett's procedure

Dunnett (1955) developed a multiple comparison procedure based on Student's t distribution for the experimental situations in which one control group is compared against each treatment group. When Dunnett's procedure is applied, a significant test of overall null hypothesis is not required because the comparisons are considered to be a priori.

Dunnett's statistic for testing the significance of the differences between a control group mean and each treatment group mean is:

$$t = \frac{\bar{Y}_i - \bar{Y}_o}{\sqrt{MSe\left(\frac{1}{n_i} + \frac{1}{n_o}\right)}}, \quad i=1, \ldots, k,$$

where $\bar{Y}_o$ represents the control group mean and $\bar{Y}_1, \ldots, \bar{Y}_k$ represent k treatment group means. The above statistic is distributed as the Student's t with ν degrees of freedom for MSe. However, an experimenter should refer to the special table developed by Dunnett (1955, 1964) for appropriate critical values. The table contains the critical values for a set of k pairs of means, i.e., between the control and k treatment groups. Dunnett's test can control the type I error rate in the sense that the probability of falsely rejecting null hypotheses is at most α for a set of k comparisons (Miller, 1981). As the number of treatment groups increases, the critical values become larger in order to keep experimentwise error rate at α (Klockars & Sax, 1986).

Dunnett (1955) has presented four tables for one- and two-sided comparisons with two levels of significance, .01 and .05 for equal n. Unfortunately, tables are not available for unequal number of observations. Dunnett suggested that the tables given for equal n may be used for unequal n as an approximation.

The four procedures, Dunnett's, CB, Scheffé's, and the regular t-test, are all based on Student's t-statistic. But each of them will be compared against a different critical value. When there are only two groups, one control and one treatment, a sample t-value and its critical value are exactly the same for all t-distribution based procedures. However, if one control is compared with more than one treatment group, the critical value required for Dunnett's test is smaller than the CB, Tukey's or Scheffé's procedure (Dunnett, 1955; Miller, 1981).

Sequentially rejective Bonferroni test

Holm (1979) has introduced a multiple comparison procedure based on the Bonferroni inequality by modifying the CB, called the sequentially rejective Bonferroni test (SRB), to increase power of the test.

1. Simple sequentially rejective Bonferroni test

When n hypotheses are tested at α simultaneously, n probability values (p-values) can be obtained. Let $P_1, P_2, \ldots, P_n$ be the p-values which are rank ordered by size from the smallest to the largest. The hypotheses are rearranged in accordance with the order of the p-values. Let the reordered hypotheses be denoted by $H_{0(1)}, H_{0(2)}, \ldots, H_{0(n)}$. It is assumed that $H_{0(1)}, H_{0(2)}, \ldots, H_{0(n)}$ are a minimal set of hypotheses, that is, any hypothesis which is the intersection of any other hypotheses is excluded from the set.

Hypotheses in the simple sequentially rejective Bonferroni procedure (simple SRB) are tested at a successively higher level of significance, i.e., a more relaxed α level. The simple SRB starts testing the first hypothesis which has the smallest p-value, P_1. The P_1 is compared with α/n, where α is the specified level of experimentwise significance and n is the number of hypotheses in the experiment. If the first hypothesis is rejected at α/n, the second hypothesis will be tested at $\alpha/(n-1)$. In general a hypothesis is tested at α/k, where k is reduced by one at each step, that is, k=n, n-1, n-2, ... , 2, 1. The sequential procedure is continued until one hypothesis is accepted or the last hypothesis, $H_{o(n),}$, is rejected. If $P_i > \frac{\alpha}{n-i+1}$, the SRB stops at Step i. Furthermore, the i^{th} hypothesis, $H_{0(i)}$, and the remaining hypotheses, $H_{o(i+1)}, H_{o(i+2)}, \ldots, H_{o(n)}$, are accepted while $H_{o(1)}, H_{o(2)}, \ldots, H_{o(i-1)}$ are rejected.

Holm (1979) has proved that in the SRB the probability of rejecting one or more true null hypotheses for free combinations, that is, any subset of the n hypotheses where n

hypotheses and the hypotheses in any subset are minimal, will be equal to or less than α.

Holm has also confirmed that his procedure is coherent and consonant according to the description given by Gabriel (1969). If a hypothesis is rejected and any other hypotheses logically implied by it are accepted, no coherence exists for such a test procedure. If a hypothesis is rejected and other hypotheses logically implied by it result in non-rejection rather than in acceptance, the test procedure contains dissonance. Coherence avoids the contradiction of rejection and acceptance among a set of minimal hypotheses, while consonance prevents the conflict of rejection and non-rejection among them. It should be noted that acceptance is not the same conclusion as non-rejection.

Holm has shown that the SRB is substantially more powerful than the CB because the level of significance used in Holm's procedure, $\frac{\alpha}{n-i+1}$, is consistently higher than that in the CB, α/n, except when i=1, where the hypothesis is tested at the same probability level by both the CB and SRB procedures. Consequently, the same or more true null hypotheses are rejected in the SRB than in the CB procedure. The power advantage of Holm's procedure (SRB) over the CB is negligible if most of the hypotheses are retained; and the gain is substantial if the number of "to be rejected" hypotheses is large.

Although Holm's procedure tests hypotheses one at a time, the simultaneous confidence intervals of the SRB will not be any shorter than that of the CB procedure. In fact, the confidence intervals constructed with decreasing critical values are not the simultaneous confidence intervals. In order to set simultaneous confidence intervals, one should use the largest critical value for all hypotheses, which is the same as in the CB procedure. Thus, the simultaneous confidence intervals in the SRB have the same length as in the CB procedure.

Flexibility is one of the greatest advantages of the SRB procedure. The SRB can be applied to any experimental situations where the CB method can be used. It requires a little more calculation than the CB but has a considerable power gain over the CB. The SRB can

be applied to multiple comparisons using both parametric and non-parametric statistics. Furthermore the SRB can be applied to one- or two-sided hypotheses; and to planned and post hoc comparisons as long as the total number of comparisons to be made is known. Although the critical values corresponding to the precise α values are not readily available, the SRB can be easily used with Bailey's (1977) extended t tables.

In his paper, Holm has pointed out the power advantage of his procedure. If six of the ten hypotheses are rejected at α=.05, the power gain in the simple SRB over the CB is about 22.6%. When the simple SRB is compared with the classical Dunnett test (Dunnett, 1955) and with the refined Dunnett's procedure (Marcus, Peritz, and Gabriel; 1976), the simple SRB is slightly more powerful than the classical Dunnett but is slightly less powerful than the refined Dunnett. However, for unequal sample sizes the simple SRB can be easily applied but the other two require complex computations because the tables for the critical values are not available for unequal number of observations. In this regard Holm's procedure is more practical for unequal n than the Dunnett test or the modified Dunnett test.

The sequentially rejective Bonferroni procedure can be generalized to the situation where different weights corresponding to the different levels of importance for hypotheses need to be taken into consideration. The method for dealing with contrasts ordered by their probability levels alone is called the simple SRB and the method for testing hypotheses with differential weights in addition to the probability levels is called the generalized SRB.

2. Generalized sequentially rejective Bonferroni test

In the generalized SRB, each p-value associated with the hypothesis, q_i, is divided by the corresponding weight, b_i, the relative importance of the hypothesis in question in the pool of n hypotheses. The ratios, q_i/b_i, thus obtained are rank ordered from the smallest to the largest. Let $s_1, s_2, ..., s_n$ be the rank ordered ratios. For more important hypotheses the larger values of weights are assigned so that the hypotheses are tested at a more relaxed or

liberal α level, which in turn, increases the power for these tests. For less important hypotheses, the reverse is true. The smaller weights assigned to these hypotheses assure that they are tested at more conservative α level.

The hypotheses and a set of different weights need to be rearranged according to rank ordered ratios, s_i. Let $H_{o(1)}, H_{o(2)}, \ldots, H_{o(n)}$ and $a_{(1)}, a_{(2)}, \ldots, a_{(n)}$ be represented by reordered hypotheses and differential weights, respectively. Note, however, $H_{o(1)}$ may not be necessarily the least important hypothesis and $H_{o(n)}$ may not the most important one, because the ratio, s_i, is a function of both the p-value and the importance weight.

In the generalized SRB a set of hypotheses are tested at successively larger α levels as in the simple SRB. The first hypothesis that has the smallest ratio, s_1, is tested against the smallest type I error rate, which is α divided by the sum of total weights, $\frac{\alpha}{\sum_{i=1}^{n} a_{(i)}}$. Given that the first hypothesis is rejected, the second hypothesis will be tested at $\frac{\alpha}{\sum_{i=2}^{n} a_{(i)}}$, where the denominator indicates the sum of weights for the second through the last hypotheses. Generally, the denominator portion of the level of significance to be applied is reduced by the weight of the currently rejected hypothesis.

The power of the generalized SRB increases for hypotheses with larger weights because the hypotheses will be tested at a more relaxed level of significance (i.e., larger p value) than the CB method. However, the power decreases for hypotheses having smaller weights. If all the weights are the same, hence, no importance of ranking is made, the generalized SRB is reduced to the simple SRB. In practice it might be difficult to quantify the importance of each hypothesis by assigning a weight unless the experimenter has a thorough knowledge of the content area.

3. Modified sequentially rejective test

According to Holm a logical implication among the hypotheses exists when the rejection (or truth) of some hypotheses implies the rejection (or truth) of some other hypotheses in a set. In applying the SRB to experimental situations, there are many different types of logical implications. One can find out which type of logical implications exists and avoid unnecessarily losing power.

Shaffer (1986) has modified Holm's procedure by considering logical implications. In general the procedure for testing hypotheses in the MSRB1 is the same as in Holm's simple SRB except for adopting different levels of significance. In the simple SRB a level of significance is $\frac{\alpha}{n-i+1}$, i=1, 2, ... , n. In the MSRB1 a level of significance $\left[\frac{\alpha}{n-i+1}\right]$ is replaced by α/t_i , where t_i is the maximum possible number of true null hypotheses. The logical implication is used in determining the value of t_i. Shaffer has described a method for determining t_i, but the procedure is complex.

Shaffer's modified sequentially rejective Bonferroni procedure (MSRB1) has more power than Holm's procedure but has limited applications. One of the situations that the logical implication can be applied is when conducting all possible pairwise comparisons among means. Shaffer has provided a table for the set of maximum number of true hypotheses, t_i, for the number of means equal to or greater than three and less than or equal to ten ($3 \leq k \leq 10$), which can be used only to test all pairwise comparisons (Shaffer, 1986, p. 828). If the number of groups are greater than ten, the recursion formula introduced by Shaffer (1986, p. 827) can be used to obtain a set of t_i.

Seaman, Franke, Serlin, and Levin (1989) have provided an example of how to obtain the number of true hypotheses, t_i, among four means. They have set a hypothetical order of mean differences among six pairs of means. Suppose that all six mean differences, μ_{D1}, μ_{D2}, μ_{D3}, μ_{D4}, μ_{D5}, μ_{D6}, are rank ordered from the largest to the smallest and that

each mean difference is obtained from the specified pair given in Table 1 according to the order set by Seaman et al. (1989). When the first hypothesis, $H_{o(1)}: \mu_1 = \mu_4$, is tested, t_i is equal to six because no pair of means is rejected yet, thus there is no partition among the four means. All the means in the same partition have no significant differences among them or not yet shown to have differences. If the present hypothesis is rejected, there are different partitions possible among means but a partition that produces the maximum possible number of true hypotheses for testing subsequent hypotheses will be sought. Among the means belonging to different partitions are significantly different. Given that the assumption of all six hypotheses are rejected in this particular sequence, Table 1 shows how to determine t_i for each step.

If no logical implication is considered, as in the simple SRB, α/k is used as the level of significance for testing the six hypotheses, where k =6 for the first hypothesis and k = 1 for the last hypothesis. But if the logical implication is taken into consideration, as in the MSRB1, t_i = 6, 3, 3, 3, 1, 1, as shown in Table 1. Because the allocated α level in the MSRB1 is equal to (at Step 1) or smaller than using the SRB, the MSRB1 is at least as powerful as the simple SRB. As the number of rejected hypotheses increases, the power advantage of Shaffer's procedure over Holm's procedure increases.

Further modification to the Shaffer's is proposed by Holland and Copenhaver (1987). Their modification (MSRB2) gives a further increase in power over Shaffer's procedure under the assumption of positive orthant dependence. A test statistic, Y_i, is positively orthant dependent if the following inequality exists for a set of test statistics, Y_i, i=1, 2, ... , n and for a set of corresponding critical values, y_i, i=1, 2, ... , n.

$$P(Y_i \le y_i;\ i=1, \dots, n) \ge \prod_{i=1}^{p} P(Y_i \le y_i) \text{ for all } y_i,\ i=1, \dots, n.$$

Holland and Copenhaver (1987) have proposed that at Step i, the significance level adopted by the MSRB1, α/t_i, be replaced by $1-(1-\alpha)^{1/t_i}$, where t_i is the maximum

number of true hypotheses as in Shaffer (1986). $1-(1-\alpha)^{1/t_i} \geq \alpha/t_i$ is always satisfied. Consequently, the MSRB2 is even more powerful than the MSRB1. The experimentwise type I error rate of the MSRB2 is controlled at or less than α. The MSRB2 has even more limited application than the MSRB1, because the MSRB2 requires both logical implications and the orthant dependence in test statistics. The improvement in power using the MSRB2 over MSRB1 is consistent but is not substantial. If both the logical implication and the positive orthant dependence are not employed or cannot be employed, the three procedures, Holm's, Shaffer's, and Holland-Copenhaver's, are all equivalent.

Holland and Copenhaver (1988) have investigated the power of the SRB and its modifications (MSRB1 and MSRB2) in factorial designs involving tests for main effects, interactions, and simple main effects. The data contained twelve individual tests--eight of the tests were obtained from chi-square tests and the other four from one-tailed t tests. The twelve p-values obtained were rank ordered from the smallest to the largest and the corresponding hypotheses were tested one at a time. The CB was compared with the simple SRB, MSRB1, and MSRB2 using the actual experimental data in a four-way ANOVA. The first eight hypotheses were rejected by all four procedures. However, the ninth hypothesis was retained by the CB and rejected by the SRB, MSRB1 and MSRB2. In a two-way ANOVA, the MSRB1 and MSRB2 were applied for the tests of simple main effects. By considering the logical implication the maximum number of true null hypotheses were determined and the familywise error was set at α=.10. Only the first and second hypotheses were rejected among the twelve and they were rejected by both the MSRB1 and the MSRB2.

In conclusion the SRB introduced by Holm (1979) is flexible and easy to use in various research situations with more power than the CB procedure. The MSRB1 by Shaffer (1986) and the MSRB2 by Holland and Copenhaver (1987) give more power than the simple SRB. The modified SRB procedures have restricted applications and the power advantage of the

MSRB1 or MSRB2 over the SRB in an actual data set was minimum (Holland & Copenhaver, 1988).

Usually, Tukey's HSD is recommended for testing all pairs of means because it is more powerful than other conventional procedures while maintaining the type I error rates at the nominal level of significance (Carmer & Swanson, 1973; Dunn, 1961; Keselman & Toothaker, 1974; Petrinovich & Hardyck, 1969; Scheffé, 1953; Seaman et al., 1989; Spjotvoll & Stoline, 1973; Ury, 1976; Ury & Wiggins, 1975).

Seaman et al. (1989) compared the Tukey test with the SRB and MSRB1 in terms of the experimentwise type I error and the power rates. For the complete-null cases, the experimentwise type I error rates of the Tukey test were close to or slightly over the nominal level of .05. When most of the mean differences were retained, the power gain of the MSRB1 over the Tukey test was negligible. However, when the most of the mean differences were rejected, the power gain of the MSRB1 over the Tukey was substantial. Although the MSRB1 is more powerful procedure than Tukey's procedure in testing all pairwise comparisons, the former procedure requires a complicated procedure to determine the maximum number of true hypotheses. The Tukey test and the SRB were slightly less powerful than the MSRB1, but the former (the Tukey procedure and the SRB) provided computational simplicity with reasonable power while keeping the familywise error rate below its nominal value. Therefore, if a researcher prefers a simple procedure with reasonable power in conducting all pairwise comparisons, either the Tukey test or the SRB seems to be the best choice.

All the procedures described above, the CB, Tukey, Scheffé, Dunnett, SRB, MSRB1 and MSRB2, control the experimentwise error rates at the nominal level of significance. The choice among these seven procedures would depend on the experimenter's decision regarding which type of contrasts is of the primary interest, which procedure is most powerful, and which procedure has computational simplicity.

Geometric presentation

In addition to the conventional procedures (CB, Tukey, Scheffé, Dunnett) for which geometric presentations are already provided (Games, 1971; Miller, 1981), some of the recent multiple comparison procedures, the SRB (Holm, 1979), MSRB1 (Shaffer, 1986), and MSRB2 (Holland & Copenhaver, 1987), are also presented geometrically in this section to aid visual understanding of these complex methods.

The maximum number of orthogonal dimensional space is the degrees of freedom associated with k means, k-1. The same set of three treatment means is used throughout this section. With three means the multiple comparison procedures can be presented in two dimensional orthogonal space. Although it is possible to use three dimensional unorthogonal space to present the retention and rejection area for the three pairwise comparisons, it is much easier to understand these three comparisons in two dimensional orthogonal space. The three pairwise comparisons are presented for each of the six procedures and the two pairwise comparisons, where each of two treatments is compared against one control, are presented for the Dunnett test. Therefore, the retention region for the Dunnett test is not directly comparable to the retention regions based on the three pairwise comparisons.

A set of orthogonal contrasts among three means are set as:

	T1	T2	T3
Ψ_1	1	0	-1
Ψ_2	1	-2	1.

All pairwise comparisons among three means are:

	T1	T2	T3
Ψ_1	1	0	-1
Ψ_3	0	1	-1
Ψ_4	1	-1	0.

The unbiased estimates of Ψ_j, j=1, 2, 3, and 4, are:

$$\hat{\Psi}_1 = \bar{X}_1 - \bar{X}_3,$$
$$\hat{\Psi}_2 = \bar{X}_1 - 2\bar{X}_2 + \bar{X}_3,$$
$$\hat{\Psi}_3 = \bar{X}_2 - \bar{X}_3,$$
$$\hat{\Psi}_4 = \bar{X}_1 - \bar{X}_2.$$

Using $\hat{\Psi}_1$ and $\hat{\Psi}_2$, the $\hat{\Psi}_3$ and $\hat{\Psi}_4$ can be reexpressed as follows:

$$\hat{\Psi}_3 = \frac{1}{2}(\hat{\Psi}_1 - \hat{\Psi}_2),$$
$$\hat{\Psi}_4 = \frac{1}{2}(\hat{\Psi}_1 + \hat{\Psi}_2).$$

The corresponding null hypotheses of the above all pairwise comparisons are:

$$H_{o1}: \mu_1 - \mu_3 = 0,$$
$$H_{o3}: \mu_2 - \mu_3 = 0,$$
$$H_{o4}: \mu_1 - \mu_2 = 0.$$

The retention area of H_{o1} is $-cv \times s_{\hat{\psi}_1} \le \hat{\psi}_1 \le cv \times s_{\hat{\psi}_1}$, where cv is the corresponding critical value and $s_{\hat{\psi}_1}$ is the standard error of first contrast depending on the procedure. The center of the retention area of H_{o1} is 0.0 and the retention region will be extended by the distance $cv \times s_{\hat{\psi}_1}$ in both directions along the axis, Ψ_1.

The following data will be consistently used to present the three pairwise comparisons of all multiple comparison procedures geometrically under the null hypotheses of no treatment differences: the number of treatment, k, is 3, dfe=30 (11 equal observations per treatment), MSe=5.5, and α=.05.

It is assumed that the means of Treatment 1 and Treatment 3 are the largest and the smallest, respectively, and the order of three mean differences from the largest to the smallest is $\hat{\psi}_1 (\bar{X}_1 - \bar{X}_3)$, $\hat{\psi}_4 (\bar{X}_1 - \bar{X}_2)$, and $\hat{\psi}_3 (\bar{X}_2 - \bar{X}_3)$.

For each contrast the region of retention for the Tukey HSD can be represented as:

$$\left|\hat{\psi}_j\right| \le q_{.\alpha;k,\nu}\sqrt{\frac{MSe}{n}},\ j=1,\ 3,\ \text{and}\ 4.$$

$$\left|\hat{\psi}_j\right| \le q_{.05;3,30}\sqrt{\frac{5.5}{11}} = 3.49\sqrt{\frac{1}{2}} = 2.4678,\ j=1,\ 3,\ \text{and}\ 4.$$

The 95% retention area of the HSD for all three pairwise comparisons on two dimensional orthogonal space (Table 2 and Figure 1).

The standard error of the j^{th} contrast in the CB, Scheffé's, Dunnett's test, SRB, MSRB1, and MSRB2 can be calculated as follows:

$$s_{\hat{\psi}_j} = \sqrt{MSe\sum_{i=1}^{k}\frac{c_i^2}{n_i}} = \sqrt{5.5\left(\frac{1^2}{11} + \frac{(-1)^2}{11}\right)} = \sqrt{5.5\left(\frac{2}{11}\right)} = \sqrt{\frac{11}{11}} = 1.$$

Because the standard error of each contrast is set to 1, the region of retention for all six procedures can be determined by the corresponding critical value alone. Tables 3 through 8 show the 95% retention areas for three pairwise contrasts for the six procedures. Each of the retention regions is geometrically presented in Figures 2 through 7 using the two dimensional orthogonal space, $\hat{\psi}_1$ and $\hat{\psi}_2$.

As the retention area becomes larger, the method becomes more conservative. As can be seen in the figures presented, the Scheffé procedure has the largest retention area among the six comparable procedures.

References

Bailey, B. J. R. (1977). Tables of the Bonferroni t statistic. Journal of the American Statistical Association, 72, 469-478.

Carmer, S. G., & Swanson, M. R. (1973). An evaluation of ten pairwise multiple comparison procedures by Monte Carlo methods. Journal of the American Statistical Association, 68, 66-74.

Dunn, O. J. (1961). Multiple comparisons among means. American Statistical Association Journal, 56, 52-64.

Dunnett, C. W. (1955). A multiple comparison procedure for comparing several treatments with a control. Journal of the American Statistical Association, 50, 1096-1121.

Dunnett, C. W. (1964). New tables for multiple comparisons with a control. Biometrics, 20, 482-491.

Gabriel, K. R. (1969). Simultaneous test procedures-some theory of multiple comparisons. The Annals od Mathematical Statistics, 40, 224-250.

Games, P. A. (1971). Multiple comparisons of Means. American Educational Research Journal, 8, 531-565.

Holland, B. S., & Copenhaver, M. D. (1987). An improved sequentially rejective Bonferroni test procedure. Biometrics, 43, 417-423.

Holland, B. S., & Copenhaver, M. D. (1988). Improved Bonferroni-type multiple testing procedure. Psycholoical Bulletin, 104, 145-149.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. Scandinavian Journal of Statistics, 6, 65-70.

Hopkins, K. D., & Chadbourn, R. A. (1967). A schema for proper utilization of multiple comparisons in research and a case study. American Educational Research Journal, 4, 407-412.

Howell, J. F., & Games, P. A. (1973). The robustness of the analysis of variance and the Tukey WSD test under various patterns of heterogeneous variances. Journal of Experimental Design, 41, 33-37.

Jaccard, J., Becker, M. A., & Wood, G. (1984). Pairwise multiple comparisons: A review. Psychological Bulletin, 96, 589-596.

Keselman, H. J., & Toothaker, L. E. (1974). Comparison of Tukey's T-method and Scheffe's S-method for various numbers of all possible differences of averages contrasts under violation of assumptions. Educational and Psychological Measurement, 34, 511-519.

Keuls, M. (1952). The use of studentized range in connection with an analysis of variance. Euphytica, 1, 112-122.

Kirk, R. E. (1982). Experimental design: Procedures for the behavioralsciences (2nd Ed.). Belmont, CA: Brooks/Cole.

Klockars, A. J., & Sax, G. (1986). Multiple comparisons. CA:, Sage Publications.

Marcus, R., Peritz, E., & Gabriel, K. R. (1976). On closed testing procedures with special reference to ordered analysis of variance. Biometrika, 63, 655-660.

Miller, R. G., Jr. (1981). Simiultaneous statistical inference (2nd Ed.). New York: Springer-Verlag.

Newman, D. (1939). The distribution of the range in samples from a normal population in terms of an independent estimate of the standard deviation. Biometrika, 31, 20-30.

Petrinovich, L. F., & Hardyck, C. D. (1969). Error rates for multiple comparison methods: Some evidence concerning the frequency of erroneous conclusions. Psychological Bulletin, 71, 43-54.

Scheffé, H. (1953). A method for judging all contrasts in the analysis of variance. Biometrika, 40, 87-104.

Scheffé, H. (1959). The analysis of variance. New Yirk: John Wiley. Inc.,

Seaman, M. A., Franke, T. M., Serlin, R. C., & Levin, J. R. (1989). New, improved multiple comparisons procedures: More pep with each step. A paper presented at the annual meeting of the American Educational Research Association in San Francisco (March, 1989).

Shaffer, J. P. (1986). Modified sequentially rejective multiple test procedures. Journal of the American Statistical Association, 81, 826-831.

Spjotvoll, E. & Stoline, M. R. (1973).An extension of the T-method of multiple comparison to include the cases with unequal sample sizes. Journal of the American Statistical Association, 68, 975-978.

Ury, H. K., & Wiggins, A. D. (1975). A comparison of three procedures for multiple comparisons among means. British Journal of Mathematical and Statistical Psychology, 28, 88-102.

Ury, H. K. (1976). A comparison of four procedures for multiple comparisons among means (pairwise contrasts) for arbitrary sample sizes. Technometrics, 18, 89-97.

Table 1

Determining maximum number of true hypotheses in testing all pairwise comparisons among four means by considering a logical implication

Step	Mean difference	p-value	Hypothesis	Partition of four means	Maximum number of true hypotheses
1	$\mu_{D1} : \mu_1 = \mu_4$	p_1	$H_{0(1)} : \mu_1 = \mu_4$	(1, 2, 3, 4)	$C_2^4 = 6$
2	$\mu_{D2} : \mu_1 = \mu_3$	p_2	$H_{0(2)} : \mu_1 = \mu_3$	(1)(2, 3, 4) or (4)(1, 2, 3)	$C_2^3 = 3$
3	$\mu_{D3} : \mu_1 = \mu_2$	p_3	$H_{0(3)} : \mu_1 = \mu_2$	(1)(2, 3, 4)	$C_2^3 = 3$
4	$\mu_{D4} : \mu_2 = \mu_4$	p_4	$H_{0(4)} : \mu_2 = \mu_4$	(1)(2, 3, 4)	$C_2^3 = 3$
5	$\mu_{D5} : \mu_2 = \mu_3$	p_5	$H_{0(5)} : \mu_2 = \mu_3$	(1)(2)(3, 4)	$C_2^2 = 1$
6	$\mu_{D6} : \mu_3 = \mu_4$	p_6	$H_{0(6)} : \mu_3 = \mu_4$	(1)(2)(3, 4)	$C_2^2 = 1$

Note: The hypothetical order of mean differences between two means from the largest to the smallest is given by Seaman at el. (1989).

Table 2

Retention Areas of the Three Pairwise Contrasts for Tukey's HSD.

Contrast	Retention area
$\hat{\Psi}_1$	$-2.4678 \le \hat{\Psi}_1 \le +2.4678$
$\hat{\Psi}_3$	$-2.4678 \le \hat{\Psi}_3 \le +2.4678 \rightarrow \hat{\Psi}_1 - 4.9356 \le \hat{\Psi}_2 \le \hat{\Psi}_1 + 4.9356$
$\hat{\Psi}_4$	$-2.4678 \le \hat{\Psi}_4 \le +2.4678 \rightarrow -\hat{\Psi}_1 - 4.9356 \le \hat{\Psi}_2 \le -\hat{\Psi}_1 + 4.9356$

Table 3

Retention Areas of the Three Pairwise Contrasts for the CB.

Contrast	Retention area
$\hat{\Psi}_1$	$-2.54 \leq \hat{\Psi}_1 \leq +2.54.$
$\hat{\Psi}_3$	$-2.54 \leq \hat{\Psi}_3 \leq +2.54 \rightarrow \hat{\Psi}_1 - 5.08 \leq \hat{\Psi}_2 \leq \hat{\Psi}_1 + 5.08$
$\hat{\Psi}_4$	$-2.54 \leq \hat{\Psi}_4 \leq +2.54 \rightarrow -\hat{\Psi}_1 - 5.08 \leq \hat{\Psi}_2 \leq -\hat{\Psi}_1 + 5.08$

Table 4

Retention Areas of the Three Pairwise Contrasts for the Scheffe's Procedure.

Contrast	Retention area
$\hat{\Psi}_1$	$-2.5768 \leq \hat{\Psi}_1 \leq +2.5768$
$\hat{\Psi}_3$	$-2.5768 \leq \hat{\Psi}_3 \leq +2.5768 \rightarrow \hat{\Psi}_1 - 5.1536 \leq \hat{\Psi}_2 \leq \hat{\Psi}_1 + 5.1536.$
$\hat{\Psi}_4$	$-2.5768 \leq \hat{\Psi}_4 \leq +2.5768 \rightarrow -\hat{\Psi}_1 - 5.1536 \leq \hat{\Psi}_2 \leq -\hat{\Psi}_1 + 5.1536.$

Table 5

Retention Areas of the Two Pairwise Contrasts for Dunnett's procedure

Contrast	Retention area
$\hat{\Psi}_1$	$-2.32 \leq \hat{\Psi}_1 \leq 2.32$
$\hat{\Psi}_3$	$-2.32 \leq \hat{\Psi}_3 \leq 2.32 \rightarrow \hat{\Psi}_1 - 4.64 \leq \hat{\Psi}_2 \leq \hat{\Psi}_1 + 4.64$

Table 6

Retention Areas of the Three Pairwise Contrasts for SRB

Contrast	level of significance	CV	Retention area
$\hat{\Psi}_1$	.05/3	2.54	$-2.54 \leq \hat{\Psi}_1 \leq +2.54.$
$\hat{\Psi}_4$	.05/2	2.36	$-2.36 \leq \hat{\Psi}_4 \leq +2.36$
			$\Rightarrow -\hat{\Psi}_1 - 4.72 \leq \hat{\Psi}_2 \leq -\hat{\Psi}_1 + 4.72.$
$\hat{\Psi}_3$	.05/1	2.042	$-2.042 \leq \hat{\Psi}_3 \leq +2.042$
			$\Rightarrow \hat{\Psi}_1 - 4.084 \leq \hat{\Psi}_2 \leq \hat{\Psi}_1 + 4.084.$

Table 7

Retention Areas of the Three Pairwise Contrasts for MSRB1

Contrast	level of significance	CV	Retention area
$\hat{\Psi}_1$	.05/3	2.54	$-2.54 \leq \hat{\Psi}_1 \leq +2.54.$
$\hat{\Psi}_4$	.05/1	2.042	$-2.042 \leq \hat{\Psi}_4 \leq +2.042$
			$\Rightarrow -\hat{\Psi}_1 - 4.084 \leq \hat{\Psi}_2 \leq -\hat{\Psi}_1 + 4.084.$
$\hat{\Psi}_3$	.05/1	2.042	$-2.042 \leq \hat{\Psi}_3 \leq +2.042$
			$\Rightarrow \hat{\Psi}_1 - 4.084 \leq \hat{\Psi}_2 \leq \hat{\Psi}_1 + 4.084.$

Table 8

Retention Areas of the Three Pairwise Contrasts for MSRB2

Contrast	level of significance	CV	Retention area
$\hat{\Psi}_1$	$1-.95^{(1/3)}$	2.5284	$-2.5284 \leq \hat{\Psi}_1 \leq +2.5284$
$\hat{\Psi}_4$	$1-.95^{(1/1)}$	2.042	$-2.042 \leq \hat{\Psi}_4 \leq +2.042$
			$\Rightarrow -\hat{\Psi}_1 - 4.084 \leq \hat{\Psi}_2 \leq -\hat{\Psi}_1 + 4.084.$
$\hat{\Psi}_3$	$1-.95^{(1/1)}$	2.042	$-2.042 \leq \hat{\Psi}_3 \leq +2.042$
			$\Rightarrow \hat{\Psi}_1 - 4.084 \leq \hat{\Psi}_2 \leq \hat{\Psi}_1 + 4.084.$

Figure 1. Geometric presentation of the Tukey HSD test.

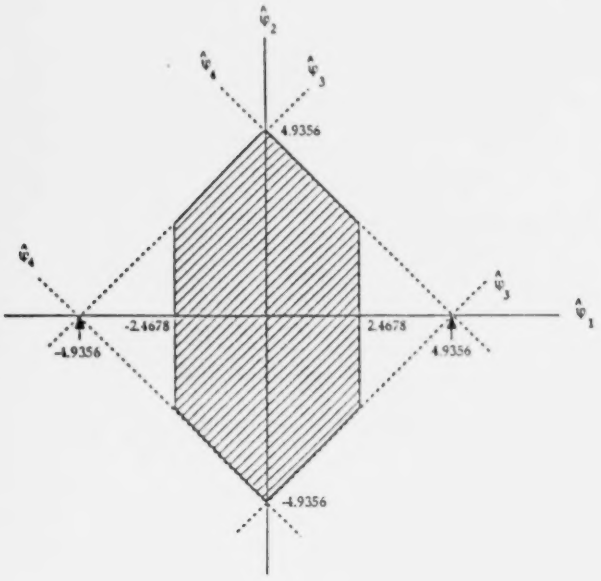

Figure 2. Geometric presentation of the CB.

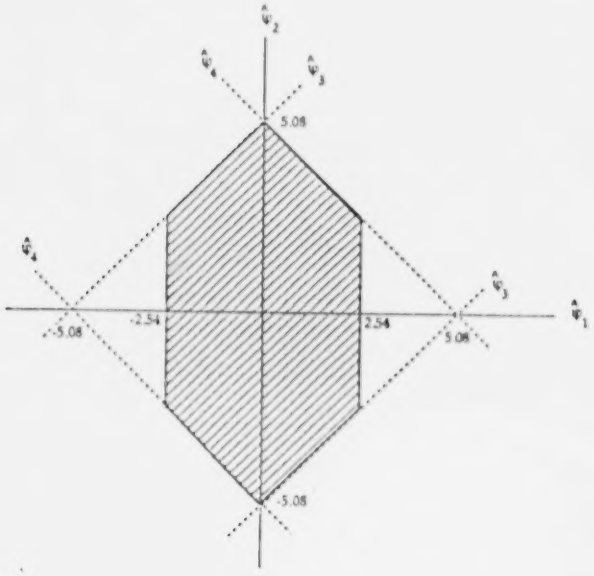

Figure 3. Geometric presentation of the Scheffe test.

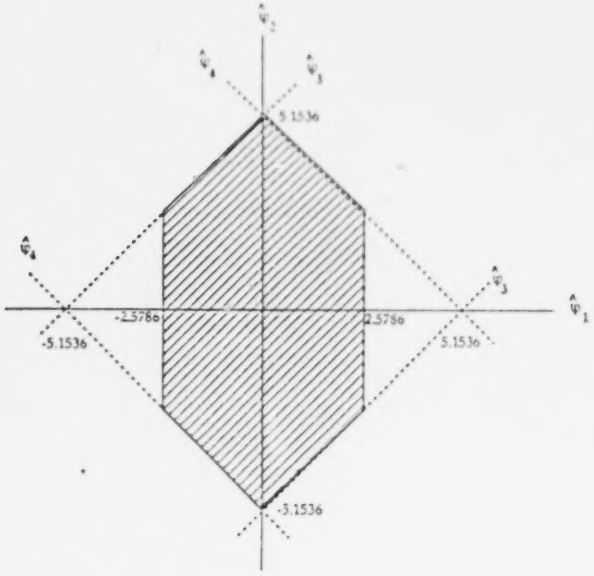

Figure 4. Geometric presentation of the Dunnett test.

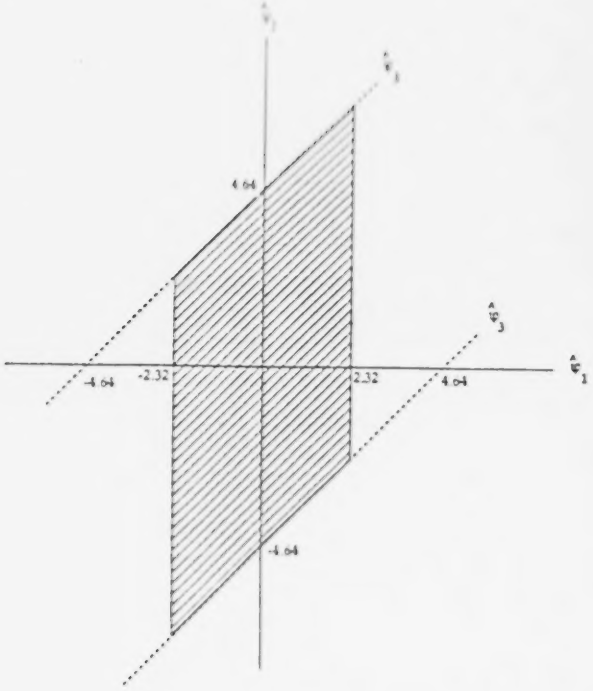

Figure 5. Geometric presentation of the SRB.

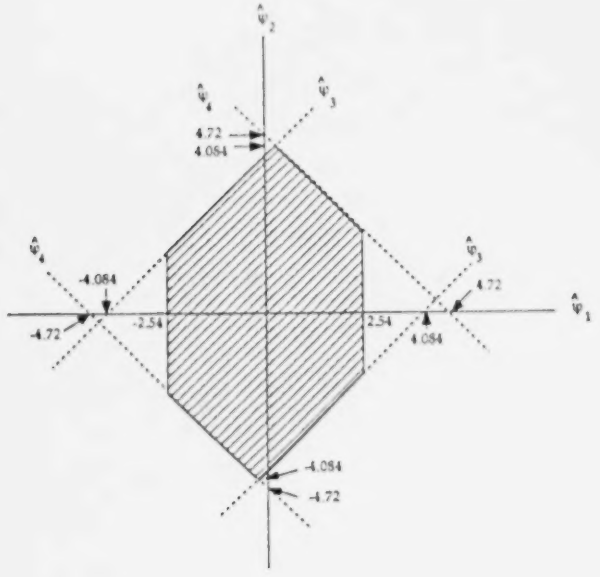

Figure 6. Geometric presentation of the MSRB1.

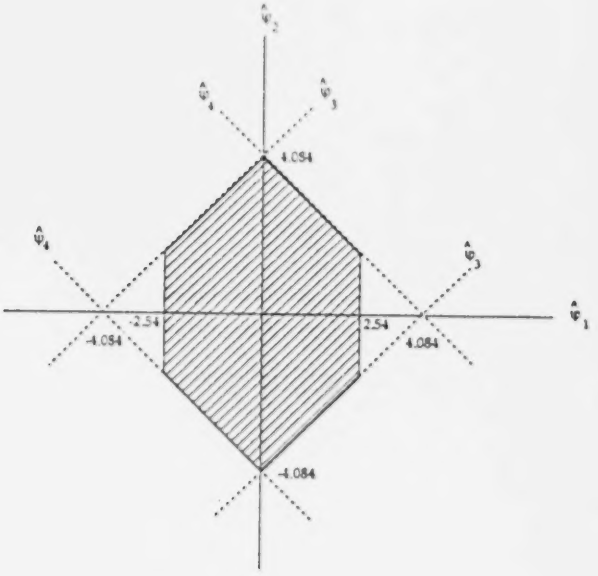

Figure 7. Geometric presentation of the MSRB2.